Online Appendix for "The Black-White Education-Scaled Test-Score Gap in Grades K-7"

Timothy N. Bond and Kevin Lang

A Theoretical Appendix and Simulations

This appendix tests some of the key assumptions of our empirical framework. Section A.1 shows that the assumption that measurement error is uncorrelated across years implies that our anchored test scores will follow a martingale, and examines the variance-covariance matrix for supportive evidence. Section A.2 discusses how small sample sizes can cause a small correlation in measurement error across time, and a simulation exercise we performed which suggests this is not a problem for our work. Section A.3 performs a robust simulation of our entire modeling environment. This simulation is primarily focused on testing our normality assumptions (linearity approximations), but also sheds light on small sample bias. Our estimator performs well, and the simulated biases would push us towards finding an increasing achievement gap, which we do not find in our main results.

A.1 Correlated Measurement Error and the Martingale Property

One concern with our approach is the assumption that measurement error is uncorrelated across grades. If this assumption is violated, at the very least our instrumental variables would be invalid. In this appendix we show that the assumption implies that our transformed test scores should evolve as a martingale. Examing the variance-covariance matrix, we find that this is not obviously violated.

First note that because θ is a forward-looking prediction,

$$\theta_{ig+1} = \theta_{ig} + \omega_{ig+1} \tag{1}$$

where ω_{ig+1} is the change in achievement between grades and

$$E\left(\theta_{ig}\Sigma_{t=1}^{T}\omega_{ig+t}\right)=0. \tag{2}$$

It then follows that last period's anchored score is also an error-ridden measure of this period's ability. Thus, θ is a martingale and s is a martingale augmented with measurement error. As discussed in Farber and Gibbons (1996), this means that the covariance of the test scores

$$\begin{aligned}\sigma_{g,g+t} &= E\left(\theta_o-\overline{\theta}_0+\Sigma_{j=1}^{g+t}\omega_j+\mu_{g+t}\right)\left(\theta_o-\overline{\theta}_0+\Sigma_{j=1}^{g}\omega_j+\mu_g\right) \\ &= \sigma_{\theta_0}^2+\Sigma_{j=1}^{g}\sigma_{\omega_j}^2+\sigma_{\mu_{g,g+t}}.\end{aligned} \tag{3}$$

The first two terms are independent of t. Under the assumption that measurement error is uncorrelated over time, the last term is 0 except when t equals 0. Note that in contrast, the covariance is increasing in g. Therefore, the model implies that the lower triangle of the covariance matrix is constant for all terms in a column below the diagonal and increasing from left to right.

Appendix table A shows the unweighted covariance matrix of the test scores.[1] We have relatively few years for which we can test this hypothesis. In all cases the covariance terms are much smaller than the variances. While we have not formally tested the hypothesis that $cov(\hat{s}_g\hat{s}_{g+2j})$ is constant for all test and grade combinations, it does not appear to be severely violated. This also suggests that the correlation in measurement error induced by some individuals sharing the same scores in tests in years g and $g-2$ is unlikely to be a serious concern, which we will address in a simulation exercise in the next section.

[1] We use $2j$ instead of j because tests are generally administered two years apart.

A.2 Small Sample Bias Simulation

In some cases more than one individual gets the same pair of scores on, for example, the first and third grade math tests. Suppose that Linda and Mike both scored in the 28th percentile in first grade and 36th percentile in third grade. Then both Linda and Mike's eventual education enter the calculation of the mean education associated with a 36 in third grade. Moreover, when we instrument for Mike's third grade education-scaled score with the mean education of everyone else with a 28 in first grade, Linda's education will also enter that calculation. This creates correlated measurement error in finite samples.[2]

To cast light on the importance of this small sample bias for our sample, we ran four simulations in which we took our actual data and added additional error to the education levels. We added a mean zero normal error with standard deviations of 1, 2, 3 and 4. Since the standard deviation of education conditional on test scores is a little less than 2 in most grades, we in effect experimented with increasing the sampling variance by 50-500 percent.

We conducted the simulation 100 times and compared the mean estimates with our actual estimates. The differences caused by this increase in the sampling error were sufficiently modest that in no case were we able to reject that the simulations produced estimates that, on average, were equal to those obtained with the actual data. And the differences between the mean simulated and actual coefficients were also visually modest, confirming that small sample bias due to correlated measurement error is not a major concern.[3] The simulation in the next section provides further evidence on this point.

A.3 Full Environment Simulation

In our empirical implementation, we estimate our measurement error correction parameter using a linear model. A linear model may not be appropriate if any of the elements used

[2] Asymptotically there will be lots of such pairs but their mean deviation from expected education will go to 0, so the IV estimator is consistent.

[3] In one case in the experiment which added N(0,16) error, there was a noticable difference between the mean estimate of the experiment and table 4, but the variance around this estimate was much too large to be meaningful.

in the estimation is not distributed normally. Therefore, in this appendix, we simulate an environment similar to that found in our data but for which we know the true evolution of the test-score gap. We then apply our approach to these simulated data to test whether the departures from normality we observe in the data lead to bias. A secondary benefit of these simulations is that we can further examine the importance of small sample bias.

We will describe our approach and results in detail shortly. In brief, we impose that the education distribution in our simulations be the distribution we observe in the data. We also impose that the distribution of test scores is uniform and takes on 99 values, corresponding to percentiles of the underlying test-score distribution. We examine three scenarios: a constant black-white test-score gap, a gap that grows each year, and a gap that grows in some but not all years. In sum, with 10,000 observations, our approach is never biased by more than 2 percent of the gap, and, in many cases, we cannot reject the null of no bias. With only 1,000 observations (somewhat fewer than most of our samples), there is evidence of modest small sample bias, but this is always less than 10 percent and averages about 6 percent of the gap. Moreover, the bias always underestimates the gap and declines as the information in the test becomes more precise. Consequently, our approach would, if anything, tend to overestimate the growth in the gap.

A.3.1 Simulation Set-up

We randomly assign half of our simulated observations to be "black." We then simulate four periods of data over which latent ability is allowed to evolve in both stochastic and predictable manners.

We begin by generating a 'true' underlying latent ability for each simulated observation. Denoting l_{it} as individual i's latent ability in period (grade) t, we assume that

$$l_{i0} = \sigma_0 \varepsilon_{i0} + b_0 B_i \tag{4}$$

where ε_{i0} is a standard normal random variable and σ_0 is a parameter chosen to calibrate our simulation to our data. Since latent ability is ordinal, the normality assumption is innocuous. The parameter b_0 is a scalar added to all black observations ($B_i = 1$) to generate an initial black-white achievement gap.

Latent ability evolves according to the formula

$$l_{it} = \kappa_t l_{it-1} + \sigma_t \varepsilon_{it} + b_t B_i \tag{5}$$

where ε_{it} is a standard normal random variable. The parameter κ_t represents growth that is predictable given prior ability levels. If $\kappa_t > 1$, prior differences in latent ability grow. The parameter σ_t determines the standard deviation of (unpredictable) changes in latent ability that are unrelated to prior latent ability. We choose κ and σ so that roughly 50 percent of the growth in the variance of latent ability in each period is due to predictable changes. Finally, b_t captures a race-specific shock that affects black observations in period t. Thus if b_t is negative, the black-white latent ability gap grows in period t.

Note that if $\sigma_t = b_t = 0$, then latent ability in periods $t-1$ and t are perfectly correlated. By both our definition and as measured in standard deviations, the true black-white gap would be unchanged between periods. By some absolute standard, it would rise or fall depending on whether κ_t is greater or less than one.

We also simulate a final schooling choice. Denote $\tilde{S}_i$ as the students final simulated latent ability when entering the labor market,

$$\tilde{S}_i = l_{i3} + \varepsilon_{is} + b_s B_i$$

where ε_{is} is a standard normal random variable. We calculate each individual's percentile in the distribution of $\tilde{S}$, and assign that individual the schooling outcome, S_i, corresponding to that percentile in the schooling distribution in the CNLSY. Thus our simulated distribution of schooling is identical to that in our data.

We simulate an underlying or latent raw test score, $\tilde{t}$, in periods 0-3 using our latent ability

$$\tilde{t}_{it} = l_{it} + \nu_{it}$$

where ν_{it} is a standard normal measurement error. While the absolute level of measurement error stays constant, the informativeness of the test varies due to changes in the simulated variance of latent ability. As we discuss below, we will calibrate our latent ability distributions in periods 0-3 to achieve a similar pattern in the simulated signal-to-noise ratio as our estimated signal-to-noise ratio for the reading recognition test in grades K-3.

We then convert $\tilde{t}_{it}$ into percentiles, so that our test scores are approximately uniformly distributed. This is consistent with our use of the percentile scores from the CNLSY.

As we do with the actual data, we then transform the t_{it} into an expected education scale, $\hat{s}_{it}$, by calculating the average level of education for everyone with the same value of t_{it} (using white observations only) as we describe in section 5 of the paper.

Estimation then proceeds as in the paper.

A.3.2 Definitions

Let, $e^*_{it} = E\left(S_i | l_{it}\right)$ which we calculate using local linear regression. The *true gap in year (or grade) t* is

$$true\ gap_t = \frac{\Sigma_{i \in white} e^*_{it}}{N_w} - \frac{\Sigma_{i \in black} e^*_{it}}{N_b},$$

or in other words, the difference between whites and blacks in average expected education given their latent ability in year t.

The *naive gap* is given by

$$naive\ gap_t = \frac{\Sigma_{i \in white} \hat{s}_{it}}{N_w} - \frac{\Sigma_{i \in black} \hat{s}_{it}}{N_b}$$

or the difference between whites and blacks in average expected education given their test score in year t.

The *IV corrected estimate* is the naive gap multiplied by our measurement error correction.

A.3.3 Choice of Parameters and Scenarios

In Table B we list the parameters we have chosen. To guide our calibration, each σ and κ was chosen to match the estimated measurement error correction from our baseline Reading Recognition tests from tables 3 and 4 under normality. When both the distribution of ability in education-scaled units and the distribution of the test measurement error is normal, this parameter $\beta = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\eta^2}$ where σ_θ^2 is the variance of the distribution of ability and σ_η^2 is the variance in the distribution of test measurement error. Since we have set $\sigma_\eta = 1$ for each test in our simulation, all variation in this parameter comes from variation in our distribution of latent ability.

With $\sigma_0 = .7239$ the signal-to-noise ratio on simulated test 0 is .34 which is roughly equivalent to .33 which we estimated for the kindergarten Reading Recognition test in the data. Adding the predicted and unpredicted period 1 changes in ability (σ_1 and κ_1) increases the variance of the white latent ability distribution to .69, and the signal-to-noise ratio of the period 1 test to .41. This compares to the Reading Recognition first grade test, where we estimate a correction parameter of .40. Likewise, in period 2 and period 3 our parameter choices create a signal-to-noise ratio of .62 and .90 on the period 2 and period 3 tests, respectively. This compares to the actual estimated parameters of .69 on the grade 2 Reading Recognition test, and .98 on the grade 3 Reading Recognition test.

We set b_0 to be roughly one half of a standard deviation of the latent ability distribution to create an additional achievement gap. The remaining b parameters increase the achievement gap by one-tenth of a standard deviation in the simulations in which they are activated.

Scenario A most closely resembles our findings. It is calibrated so that the true gap is .57 years of predicted education for all four years.

Scenario B is designed to capture the steady growth in the gap implied by the naive

estimator, albeit from a somewhat higher base than we find. Thus the gap grows steadily from .57 years of predicted schooling in "first grade" to 1.03 years in "third grade." These numbers are designed to represent a roughly one-standard deviation increase in the gap each year and between the last test and completed schooling.

Finally, to verify that our approach can capture one time growth in the gap, scenario C sets the gap at .57 years in "kindergarten" and "first grade" and at .72 years in "second" and "third" grade. We also allow for a further increase in the gap between "third grade" and the completion of schooling.

A.3.4 Results

We report the results of our simulations in Table C. Columns (1)-(3) report results from simulations with 10,000 observations (5,000 white) to demonstrate biases for large samples, while (4)-(6) report results using 1,000 observations (500 white), which is slightly less than the number of observations we have in a typical grade in the CNLSY data. We perform each simulation 500 times and report mean and standard deviations for our results.

In Panel A, we report results for scenario A (the latent achievement gap does not grow in any period). Thus our estimate of ability should be constant across all periods. In our 10,000 observation simulation, our IV estimator is almost identical to the latent ability gap in each period. In our smaller sample simulation, our estimator has a small downward bias in the early period of our simulation when the simulated test is least informative, biasing us towards finding an increasing gap. In contrast, the naive estimator has a large downward bias in high-measurement error periods. Under both the large and small sample simulations, using the naive estimator makes it appear that the black-white achievement gap is increasing.

In Panel (scenario) B, the latent ability gap increases by roughly one standard deviation each period. We should thus expect to see a steady increase in the achievement gap. Our IV estimator provides this, with almost no bias in the large sample simulation and only a small downward bias in the small sample simulation. Again, this bias is somewhat larger

when the test is less informative, biasing our estimates towards finding a somewhat larger increase in the gap than is correct. In contrast, the naive estimator greatly overstates the growth of the gap in both simulations.

In Panel (scenario) C, the latent ability gap increases by roughly one standard deviation between period 1 and period 2 and between period 3 and schooling but not between period 0 and period 1 or between periods 2 and 3. Again, our IV estimator shows little or no bias when we have 10,000 observations. With the smaller sample, we again find that there is a slight bias towards finding an increasing gap, but our estimator continues to substantially outperform the naive estimator. The naive estimates imply that the achievement gap grows in every period, not just in period 2.

B Data Appendix

B.1 Weighting Procedure

The sample is not nationally representative because children born before 1982, when the mothers were age seventeen through twenty-five, are observed only during their later childhood, while those born in later years are observed only during their early childhood. To correct for this non-representativeness, we use the CNSLY's custom weighting program to create separate weights for each grade-test designed to make that subsample nationally representative subject to the caveat that we can never observe a child born to a fourteen-year-old mother before the child turns seven or a child born to a forty-four year old mother when that child is older than nine.[4] Individuals with a valid PIAT-RR raw score below the threshold for taking the RC are included in the construction of the weights for the PIAT-RC but are excluded from the analysis. This avoids putting undue weight in the early grades on a small number of low achieving students who advance to the RC due to randomly high scores. The

[4]We are grateful to Jay Zagorsky of the Center for Human Resources Research for providing us with the program.

RC results should, therefore, be interpreted as representative of the population that would have scored sufficiently well on the RR to take the RC exam in that grade. Note that we should view gaps based on the RC with caution especially in kindergarten but also in first grade because the students taking the exam are not fully representative of the overall student population.

We also recognize that these selection issues can be even more severe for our adult sample for which we calculate the expected schooling completion for each test score. We therefore construct a second set of weights (using the CNSLY's custom weighting program) for each grade-test combination using only this sample, which we use when constructing our scale.

C Empirical Appendix

C.1 Estimates Using Full Sample to Anchor Test Scale

The results we present in the main body used only whites to anchor the test scale to educational attainment. As black student scores are disproportionately concentrated near the bottom of the distribution, this avoids the concern that our re-calibrated scores may be picking up things such as overt discrimination later in life, rather than the predictive human capital content of the test. In tables D-F, we replicate the results of Tables 3-5 uses all individuals for whom we observe an adult education outcome to anchor our scales. The results are similar to those in the main body of the paper.

C.2 High School and College Completion Anchors

In this appendix we explore using two alternative anchors based on discrete education outcomes: high school completion and college completion. We follow the same procedure as described in the main text to construct the anchor (using only whites) and present estimates that use the custom sample weights.

Table G shows the results using high school completion, while table H shows the results

for college completion. The results largely mirror those of the main specifications. The results are much less precise, but do not point to a rising achievement gap in either case. The point estimates suggest a flat or declining gap.

C.3 Unified Measure of Achievement

In the main text we used adult outcomes as a way to scale each individual subject test. In this subsection we combine information from all three tests to estimate a single measure of achievement in each grade by forming a conditional expectation of future achievement

$$E[\theta_{ig}|T_{ig}] = h(T_{ig})$$

where T is the set of tests available for student i and h is the conditional expectation function. Analogous to our earlier discussion, we do not observe achievement directly but observe eventual educational attainment, which reflects achievement in grade g. Following the theory laid out previously, if we can estimate h, we can use instrumental variables to create corrected achievement gaps for each grade.

We estimate h using a multivariate kernel Nadarya-Watson regression estimator. For a set of test scores T, the estimator creates weights for each observation based on the closeness of its test scores to T.[5] The estimator then uses these weights to form a weighted average of the outcome variable (in our case, education). Thus we can generate an expected outcome conditional on the full set of tests.

The weights depend on the choice of kernel function and bandwidth. We select a multivariate Gaussian kernel. For each point, the kernel weights observations around the point so that the density is multivariate normal. The choice of kernel is inconsequential; however the bandwidth is not (Blundell and Duncan 1998). In a multivariate Gaussian kernel, the bandwidth essentially determines the variance of the density. For bandwidth selection, we follow

[5] Given the focus of this paper, calculating how "close" test scores are is clearly problematic. We use the distance in percentile ranks. Using distance on other scales could certainly lead to other results.

Silverman's (1986) rule of thumb, so that the bandwidth is proportional to the variance of the distribution in the data.

As previously noted, many children do not advance to the reading comprehension test during the first two years of school. To account for this, we estimate the conditional expectations separately for those who did and did not take the RC exam. In the remaining grades, the very small sample of children who do not advance to the reading comprehension exam is dropped from the analysis.

Table I displays the results of this exercise. The first column shows our achievement gap estimates using only the differences in conditional expectations, not adjusting for measurement error. We see the familiar pattern of a rising initial achievement gap. Based on their performance on all tests in kindergarten, blacks are projected to obtain .41 fewer years of education than whites do. This gap rises quickly, however, to .76 years in second grade and remains roughly constant thereafter. In this respect, the results are more similar to those in Table 3 for math than for either of the reading tests. This may reflect the poor ability of the early reading tests to predict educational attainment.

However, once we correct for measurement error, the growth in the gap again disappears. The estimates in column two project a future racial difference in educational attainment of .77 years in kindergarten, with little change through seventh grade relative to the year-to-year variation. Once again, the projected education gap is at least as high if not higher than the actual education gap, consistent with Lang and Manove (2011).

The results in table I are broadly consistent with those in Table 4. In every grade except 6th, the estimated gap when we use all three tests lies within the range of the gaps produced by using each of the three tests individually. However, the confidence intervals are consistently tighter when we use all three tests, and our estimates appear meaningful even for kindergarten and first grade.

Moreover, the gap averages about .9 years of education, almost exactly what we obtain using the early PPVT test. This suggests that the difference between the results using the

PPVT and PIAT tests may not be their content but simply greater measurement error in the latter although we cannot test this directly

In table J, we repeat the exercise but instead scale the tests to represent the education-predicted mean log earnings of each score. We find similar results to those of table I. The achievement gap remains steady at about a 12 percent earnings difference, which is on par with that shown for the math achievement tests in Table 5. Our estimates are also generally more precise than in Table 5, though the improvement in precision is not nearly as substantial as with the education-scaled scores.

References

[1] Blundell, Richard and Alan Duncan. 1998. "Kernel Regression in Empirical Microeconomics." *Journal of Human Resources* 33(1):62-87.

[2] Farber, Henry S., and Robert Gibbons. 1996. "Learning and Wage Dynamics." *Quarterly Journal of Economics* 111(4):1007-47.

[3] Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Table A: Test Score Covariance Matrix

	PPVT	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
PPVT	0.83								
MATH									
Grade K	0.35	1.54							
Grade 1	0.33		1.26						
Grade 2	0.38	0.58		1.62					
Grade 3	0.36		0.60		1.46				
Grade 4	0.32	0.46		0.62		1.33			
Grade 5	0.38		0.58		0.78		1.71		
Grade 6	0.37	0.45		0.64		0.60		1.45	
Grade 7	0.37		0.55		0.74		0.80		1.44
READING RECOGNITION									
Grade K	0.21	1.33							
Grade 1	0.27		1.24						
Grade 2	0.33	0.41		1.26					
Grade 3	0.39		0.72		1.42				
Grade 4	0.29	0.40		0.67		1.55			
Grade 5	0.35		0.63		0.85		1.80		
Grade 6	0.34	0.39		0.75		0.73		1.66	
Grade 7	0.36		0.66		0.85		0.88		1.72
READING COMPREHENSION									
Grade K	0.13	1.04							
Grade 1	0.30		1.27						
Grade 2	0.34	0.23		1.29					
Grade 3	0.37		0.53		1.53				
Grade 4	0.31	0.09		0.54		1.44			
Grade 5	0.30		0.43		0.55		1.30		
Grade 6	0.40	0.20		0.56		0.61		1.64	
Grade 7	0.40		0.40		0.63		0.56		1.46

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Covariances are calculated using all available observations for each individual cell and are unweighted. Covariances of tests taken 1, 3, 5, or 7 years apart are not shown because the sample is surveyed every two years.

Table B: List of Simulation Parameters

	(1)	(2)	(3)
		Simulation	
	A	B	C
σ_0		0.7239	
b_0		-0.36	
κ_1		1.07	
σ_1		0.3425	
κ_2		1.38	
σ_2		0.7919	
κ_3		1.98	
σ_3		2.35	
b_1	0	-0.82	0
b_2	0	-0.1385	-0.1385
b_3	0	-0.3614	0
b_s	0	-0.3614	-0.3614

Notes: Parameter values for simulation model. A corresponds to no growth model. B corresponds to steady growth model. C corresponds to unsteady growth model

Table C: Simulated Estimator Performances

	(1)	(2)	(3)	(4)	(5)	(6)
	10,000 observations			1,000 observations		
	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Panel A: No Growth						
True Gap	0.57	0.57	0.57	0.56	0.57	0.57
Naive Estimate	0.24	0.38	0.53	0.24	0.38	0.53
Average Naive Miss	0.33	0.19	0.05	0.32	0.19	0.05
	(0.02)	(0.02)	(0.01)	(0.10)	(0.09)	(0.05)
IV Corrected Estimate	0.56	0.57	0.57	0.51	0.53	0.56
Average IV Miss	0.01	0.01	0.01	0.05	0.03	0.02
	(0.04)	(0.02)	(0.01)	(0.18)	(0.11)	(0.05)
Panel B: Steady Growth						
True Gap	0.70	0.84	1.03	0.70	0.84	1.02
Naive Estimate	0.28	0.55	0.93	0.29	0.54	0.93
Average Naive Miss	0.42	0.29	0.10	0.41	0.29	0.09
	(0.02)	(0.02)	(0.01)	(0.10)	(0.09)	(0.05)
IV Corrected Estimate	0.69	0.84	1.01	0.64	0.80	1.00
Average IV Miss	0.01	0.01	0.02	0.06	0.04	0.02
	(0.04)	(0.03)	(0.01)	(0.20)	(0.12)	(0.05)
Panel C: Unsteady Growth						
True Gap	0.58	0.72	0.72	0.57	0.72	0.72
Naive Estimate	0.24	0.47	0.66	0.24	0.47	0.65
Average Naive Miss	0.34	0.25	0.07	0.33	0.24	0.06
	(0.02)	(0.02)	(0.01)	(0.10)	(0.08)	(0.05)
IV Corrected Estimate	0.57	0.71	0.71	0.52	0.68	0.70
Average IV Miss	0.01	0.01	0.01	0.05	0.04	0.02
	(0.04)	(0.03)	(0.01)	(0.18)	(0.11)	(0.05)

Notes: Average estimates over 500 repititions of simulation. Standard deviations in parenthesis. Columns (1)-(3) have 10000 simulated observations (5000 whites). Columns (4)-(6) have 1000 simulated observations (500 whites). Panel A has no growth in simulated black-white achievement gap. Panel B adds 0.1 standard deviations between each test and again between test 3 and schooling. Panel C adds 0.1 standard deviations between test 1 and test 2, and between test 3 and schooling.

Table D: Raw Difference in Expected Grade Completion conditional on Test Score

	(1) Math	(2) Read-RR	(3) Read-RC
Pre-Age 5 PPVT		0.88	
		[1.12, 0.65]	
Kindergarten	0.55	0.20	0.26
	[0.33, 0.72]	[0.08, 0.36]	[-0.02, 0.52]
First Grade	0.50	0.35	0.32
	[0.37, 0.66]	[0.19, 0.46]	[0.17, 0.48]
Second Grade	0.72	0.58	0.48
	[0.56, 0.96]	[0.36, 0.77]	[0.26, 0.60]
Third Grade	0.67	0.60	0.61
	[0.52, 0.84]	[0.49, 0.78]	[0.46, 0.75]
Fourth Grade	0.70	0.56	0.58
	[0.57, 0.88]	[0.40, 0.71]	[0.44, 0.77]
Fifth Grade	0.69	0.47	0.51
	[0.54, 0.85]	[0.30, 0.61]	[0.33, 0.63]
Sixth Grade	0.70	0.58	0.60
	[0.55, 0.90]	[0.41, 0.79]	[0.40, 0.76]
Seventh Grade	0.71	0.54	0.57
	[0.54, 0.85]	[0.38, 0.70]	[0.44, 0.75]

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Point estimates represent difference between average white and average black predicted education conditional on test score for each grade-test combination. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for those who are observed age 22 or above and applied to the full sample. All results are weighted to be nationally representative.

Table E: Measurement Error Adjusted Difference in Ability in Units of Predicted Education

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	1.24	0.64	1.32
	[0.65, 2.07]	[0.17, 1.68]	[-3.04, 7.41]
First Grade	1.01	0.88	0.64
	[0.57, 1.54]	[0.38, 1.40]	[0.25, 1.15]
Second Grade	1.05	0.81	0.91
	[0.50, 1.52]	[0.35, 1.37]	[0.43, 1.48]
Third Grade	1.02	0.65	0.69
	[0.47, 1.55]	[0.41, 0.96]	[0.23, 1.09]
Fourth Grade	1.05	0.57	0.71
	[0.68, 1.56]	[0.29, 0.78]	[0.12, 1.06]
Fifth Grade	0.81	0.60	0.67
	[0.52, 1.08]	[0.32, 0.75]	[0.36, 0.87]
Sixth Grade	0.91	0.74	0.81
	[0.62, 1.18]	[0.50, 1.07]	[0.48, 1.06]
Seventh Grade	0.89	0.68	0.72
	[0.54, 1.12]	[0.43, 0.90]	[0.37, 1.19]

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Point estimates represent difference between average white and average black predicted education conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for those who are observed age 22 or above and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table F: Difference in Ability in Education-Predicted Log Earnings

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	0.17	0.09	0.15
	[0.07, 0.40]	[0.00, 0.41]	[-0.36, 0.94]
First Grade	0.12	0.11	0.08
	[0.06, 0.17]	[0.04, 0.19]	[0.03, 0.14]
Second Grade	0.13	0.12	0.12
	[0.06, 0.20]	[0.06, 0.23]	[0.05, 0.23]
Third Grade	0.12	0.09	0.09
	[0.05, 0.18]	[0.05, 0.13]	[0.03, 0.16]
Fourth Grade	0.13	0.07	0.08
	[0.07, 0.20]	[0.02, 0.10]	[-0.01, 0.13]
Fifth Grade	0.11	0.08	0.09
	[0.06, 0.15]	[0.04, 0.10]	[0.05, 0.12]
Sixth Grade	0.11	0.09	0.10
	[0.07, 0.16]	[0.06, 0.14]	[0.05, 0.13]
Seventh Grade	0.12	0.10	0.10
	[0.07 0.15]	[0.05, 0.13]	[0.05, 0.17]

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Point estimates from difference between average white and average black mean log-earnings of predicted education conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for whites who are observed age 22 or above and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use log-earnings predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table G: Measurement Error Adjusted Difference in Ability in Units of Predicted Rate of High School Completion

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	0.142	0.152	0.225
	[-0.207, 0.645]	[-0.982, 1.13]	[-2.16, 2.51]
First Grade	0.120	0.092	0.072
	[0.027, 0.516]	[0.009, 0.278]	[-0.194, 0.489]
Second Grade	0.079	0.106	0.086
	[-0.183, 0.220]	[-0.395, 0.732]	[-0.018, 0.288]
Third Grade	0.115	0.068	0.033
	[-0.236, 0.876]	[-0.127, 0.260]	[-0.431, 0.496]
Fourth Grade	0.116	0.055	0.140
	[0.033, 0.222]	[-0.036, 0.101]	[0.043, 0.296]
Fifth Grade	0.115	0.036	0.050
	[0.013, 0.236]	[-0.000, 0.060]	[-0.005, 0.143]
Sixth Grade	0.058	0.083	0.096
	[-0.110, 0.307]	[0.027, 0.161]	[0.025, 0.136]
Seventh Grade	0.076	0.050	0.063
	[0.004, 0.132]	[-0.010, 0.103]	[-0.117, 0.214]

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Point estimates represent diifference between average white and average black predicted high school completion conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for whites who are observed age 22 or above and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use predicted high school completion conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table H: Measurement Error Adjusted Difference in Ability in Units of Predicted Rate of College Completion

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	0.205	0.086	0.222
	[-0.981, 1.461]	[-0.504, 0.794]	[-0.823, 0.786]
First Grade	0.100	0.117	0.092
	[0.007, 0.202]	[-0.166, 0.516]	[-0.011, 0.222]
Second Grade	0.094	0.162	0.097
	[-0.107, 0.401]	[0.000, 0.499]	[-0.247, 0.629]
Third Grade	0.087	0.091	0.124
	[-0.154, 0.186]	[0.021, 0.161]	[-0.038, 0.366]
Fourth Grade	0.103	0.054	0.036
	[-0.128, 0.349]	[-0.117, 0.154]	[-0.453, 0.406]
Fifth Grade	0.105	0.077	0.072
	[0.018, 0.169]	[0.015, 0.123]	[-0.110, 0.137]
Sixth Grade	0.096	0.087	0.044
	[0.015, 0.171]	[0.015, 0.183]	[-0.098, 0.143]
Seventh Grade	0.108	0.106	0.110
	[-0.010, 0.182]	[0.032, 0.153]	[0.024, 0.257]

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Point estimates represent difference between average white and average black predicted college completion conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for whites who are observed age 22 or above and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use predicted college completion conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table I: Difference in Predicted White Education Using All Tests

	(1) Unadjusted	(2) IV Adjusted
Kindergarten	0.41	0.77
	[0.23, 0.56]	[0.42, 1.12]
First Grade	0.51	0.98
	[0.34, 0.69]	[0.57, 1.42]
Second Grade	0.76	0.99
	[0.56, 0.89]	[0.63, 1.29]
Third Grade	0.77	0.88
	[0.63, 0.93]	[0.68, 1.17]
Fourth Grade	0.80	0.92
	[0.64, 1.00]	[0.70, 1.19]
Fifth Grade	0.72	0.76
	[0.50, 0.82]	[0.52, 0.88]
Sixth Grade	0.81	0.95
	[0.68, 1.01]	[0.76, 1.18]
Seventh Grade	0.78	0.85
	[0.54, 0.92]	[0.59, 1.07]

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Point estimates represent difference between average white and average black predicted education for whites conditional on all test scores for each grade-test combination. Bootstrapped 95 percent confidence intervals in brackets. Column 2 estimates are corrected for measurement error by instrumental variables. Conditional predicted education computed for those who are observed age 22 or above using a multivariate kernal regression and applied to the full sample. Kindergarten and first grade use the predicted education conditional on test score for the PPVT as an instrument, while the remaining grades use that measure lagged two grades. All results are weighted to be nationally representative.

Table J: Difference in Education-Predicted White Log Income Using All Tests

	(1) Unadjusted	(2) IV Adjusted
Kindergarten	0.05	0.10
	[0.03, 0.07]	[0.05, 0.15]
First Grade	0.07	0.12
	[0.04, 0.09]	[0.07, 0.17]
Second Grade	0.10	0.12
	[0.07, 1.11]	[0.07, 0.16]
Third Grade	0.10	0.12
	[0.08, 0.12]	[0.09, 0.16]
Fourth Grade	0.10	0.11
	[0.08. 0.12]	[0.08, 0.14]
Fifth Grade	0.09	0.11
	[0.06, 0.11]	[0.07, 0.12]
Sixth Grade	0.10	0.12
	[0.08, 0.13]	[0.09, 015]
Seventh Grade	0.10	0.12
	[0.07, 0.12]	[0.08, 0.15]

Notes: Data are from the Children of the National Longitudinal Survey of Youth, 1986-2010 waves. Point estimates represent difference between average white and average black mean log-earnings of predicted white education conditional on all test scores for each grade-test combination. Bootstrapped 95 percent confidence intervals in brackets. Column 2 estimates are corrected for measurement error by instrumental variables. Conditional predicted education computed for those who are observed over 22 or above using a multivariate kernal regression and applied to the full sample. Kindergarten and first grade use the predicted education conditional on test score for the PPVT as an instrument, while the remaining grades use that measure lagged two grades. All results are weighted to be nationally representative.